Hardware requirements for big data study

spalan **2020-12-18**

As part of a research project, I will have to analyze a dataset of 8 columns and 650 million rows. I can only work on the data for about two and a half months, so I need to make sure I'll be able to run my analyses when I get the data (in about a month's time). The analyses will involve calculating effect size estimates, z/t-statistics and corresponding p-values for different hypotheses. Sorry I cannot be more specific - part of the project is not to know the exact hypotheses in advance.

My question now is what kind of hardware would be good to work on this data. I have worked with large datasets before, but none this large. I have a modern workstation with i7 processor, sufficient disk space and currently 16GB of RAM, running Windows 10 and the latest version of RStudio. I assume RAM will be the bottleneck. Does anyone have a suggestion regarding how much I should have to work with this kind of data?

Best,
Stefan.

---

nirgrahamuk **2020-12-18**

Assuming 8 columns each of 64bit numeric for 650million rows, amounts to just under 42GB

---

spalan **2020-12-18**

Sorry, I can actually concretize the column classes. There is:

1 date/time
4 integers
3 booleans

Skip to main content

I just created some sample data (using only 650000 rows) and object_size() tells me a data.table of these dimensions (scaled up by factor 1000) would require 23GB. But that would mean that if I used any operation that duplicated the data in memory, this would not work, right?

---

nirgrahamuk **2020-12-18**

Right. I would look at if I could summarise the data.
Also a lot of boot strap sampling with any given sample only being a small fraction of the total.

Either that, or pay $$$ for cloud compute

---

spalan **2020-12-18**

Right. I will probably not use all columns at all times, so that will further cut down on memory requirements. I'll check cloud computing offerings and consider my options.

Thanks for the advice!

Skip to main content

New & Unread Topics

Topic	Replies	Views	Activity
How to use `EmbedDimension()` of the rEDM package in the case of imputation of missing data in time series?	0	40	2026-05-19
C++ compile error "too many sections" only in Rstudio rstudio c	0	22	2026-06-26

Want to read more? Browse other topics in or view latest topics.

Powered by Discourse